INFORMATION TECHNOLOGY, COMPUTER SCIENCE, AND MANAGEMENT

Check for updates

UDC 004.934
https://doi.org/10.23947/2687-1653-2022-22-2-169-176

Original article

Analysis of Natural Language Processing Technology: Modern Problems and Approaches

Maria A. Kazakova, Alina P. Sultanova
Kazan National Research Technical University named after A. N. Tupolev–KAI, 10, K. Marx St., Kazan, Russian Federation
✉ kazakovamaria2609@gmail.com

Abstract

Introduction. The article presents an overview of modern neural network models for natural language processing. Research into natural language processing is of interest as the need to process large amounts of audio and text information accumulated in recent decades has increased. The most discussed in foreign literature are the features of the processing of spoken language. The aim of the work is to present modern models of neural networks in the field of oral speech processing.

Materials and Methods. Applied research on understanding spoken language is an important and far-reaching topic in the natural language processing. Listening comprehension is central to practice and presents a challenge. This study meets a method of hearing detection based on deep learning. The article briefly outlines the substantive aspects of various neural networks for speech recognition, using the main terms associated with this theory. A brief description of the main points of the transformation of neural networks into a natural language is given.

Results. A retrospective analysis of foreign and domestic literary sources was carried out alongside with a description of new methods for oral speech processing, in which neural networks were used. Information about neural networks, methods of speech recognition and synthesis is provided. The work includes the results of diverse experimental works of recent years. The article elucidates the main approaches to natural language processing and their changes over time, as well as the emergence of new technologies. The major problems currently existing in this area are considered.

Discussion and Conclusions. The analysis of the main aspects of speech recognition systems has shown that there is currently no universal system that would be self-learning, noise-resistant, recognizing continuous speech, capable of working with large dictionaries and at the same time having a low error rate.

Keywords: Natural Language Processing, oral speech, neural networks, automated natural language processing, semantic consistency.

Acknowledgements: the authors are deeply grateful to the readers for the useful comments and recommendations made during the review of the article.

For citation: M. A. Kazakova, A. P. Sultanova. Analysis of natural language processing technology: modern problems and approaches. Advanced Engineering Research, 2022, vol. 22, no. 2, pp. 169–176. https://doi.org/10.23947/2687-1653-2022-22-2-169-176

© Kazakova M. A., Sultanova A. P., 2022

Introduction. This article provides an overview of the main language model based on the neural network for Natural Language Processing that helps computers communicate with people in their native language and scale other language tasks. Modern machine learning technologies allow computers to read text, hear speech, interpret it, measure moods, and determine which parts of speech are important. This technology is called Natural Language Processing (NLP), it is based on many disciplines, including computational linguistics. NLP is increasingly being used in interactivity and productivity applications, such as creating spoken dialogue systems and speech-to-speech engines, searching social networks for health or financial information, detecting moods and emotions towards products and services, etc.

The relevance of NLP is primarily associated with the need to process large amounts of audio and text information accumulated by mankind over the past decade. Currently, most modern devices are endowed with a voice control function, and various kinds of digital assistants are becoming widespread. Now, the speech recognition function is available in almost any gadget, it allows us to interact through voice applications, facilitating and simplifying a person's life. There are a fairly large number of commercial speech recognition systems, among the most famous there are Google, Yandex, Siri. The quality of speech recognition in such systems is at a fairly high level, but they are not without a number of shortcomings. Unfortunately, despite the amazing development of computer technology, the current problem of equipping a computer with a full-fledged, natural human voice interface is still far from over.

Materials and Methods. NLP technology is rapidly advancing due to the increased interest in the field of machine learning, as well as the availability of big data, powerful computing, and improved algorithms. However, NLP is not a new science. Attempts to teach computers to communicate with people through a natural voice interface have been made since the early days of computer technology. NLP was born with the advent of the first computers from the idea of how good it would be to use these machines to solve various useful tasks related to natural language, e.g., these programs were intended for people who, due to physiological characteristics, could not type text manually [1].

The first task that the first computers solved in the early stages of the formation of NLP was the task of machine translation, i.e., automatic translation of text from one language to another using a computer. This problem was successfully solved and started to be applied in the mid-1950s, in the past century, for the “Russian-English” pair [2].

The second task of machine learning was to create conversational systems, the programs to conduct a dialogue with a person in natural language. Many systems created at that time were imperfect due to a number of difficulties in speech recognition that can have a significant impact on the quality of the result. The difference in the voices of speaking people, the inconsistency of colloquial speech, the phonogram of the same words can vary greatly depending on a number of factors: pronunciation speed, regional dialect of the language, foreign accent, social class, and even the gender of a person [3].

The third task was to create a question-and-answer system. There was a need for programs that would answer exactly the human question. At that stage, such a question was in the form of a natural language text. Thus, the problem of scaling the recognition system has always been a significant obstacle. In the course of many years of research, it has been found that it is required to involve not only programmers, but also experts in linguistics, radio engineers, mathematicians, biologists, and even psychologists in solving the problem.

At different times, various mathematical, statistical, logical, stochastic approaches were used in natural language processing, such as Dynamic Time Warping, Bayesian discrimination, Hidden Markov Model, formal grammars, and probabilistic approaches. At the present stage of natural language processing, machine learning methods are widespread, in particular, neural networks. Currently, in modern linguistic research, at the first stage, texts are selected that are planned to be analyzed, and a corpus of texts is created. Next step, the collected material is transferred to an expert linguist. He prescribes the rules, compiles dictionaries, marks up texts for the identification of target structures in

texts for the further task solution. Another method is also used, in which an expert linguist marks the text into target structures or categorizes texts in certain classes, and then machine learning methods automatically derive some rules or models for further solving current problems. At the end of the work, the quality of the methods is always checked.

Philologists study semantics of the text considering meanings of polysemantic units in context, emphasizing that context plays a fundamental role in the word definition. Therefore, e.g., the authors discover the contextual meanings of polysemantic units that are not registered in lexicographic sources. In the early stages, scientists proposed to divide any sentence into a set of words that could be processed individually, which was much easier than processing a whole sentence. This approach is similar to the one used to teach a new language to children and adults. When we first start learning a language, we are introduced to its parts of speech. Let us consider English as an example. It has 9 main parts of speech: noun, verb, adjective, adverb, pronoun, article, etc. These parts of speech help to understand the function of each word in a sentence. However, it is not enough to know the category of a word, especially for those that may have more than one meaning. Specifically, the word "leaves" can be the verb "to leave" in the 3rd person singular or the plural form of the noun "leaf", which should be considered from the point of language as a system of interrelated and interdependent units. The idea of consistency in the lexical and semantic sphere of language was first expressed by M. M. Pokrovsky, emphasizing that "words and their meanings do not live a separate life from each other, but are connected (in our soul), regardless of our consciousness, into different groups, and the basis for their grouping is similarity or direct opposition in their basic meaning". Paradigmatic, syntagmatic, and epidigmatic relations among language units are important manifestations of the systematic and regular nature of language. The researchers note that words enter the syntagmatic relations based on the logical contiguity of concepts and, consequently, their compatibility with each other [4].

We need to understand that from the point of view of computer science, speech is not structured information, but a sequence of characters. To ensure that voice data can continue to be used, the speech recognition application translates it into text. The accent, individual intonations, and emotions are already being erased in the text. When data are translated into text, they are translated with zeros and ones.

Therefore, computers need a basic understanding of grammar to refer to it in case of confusion. Thus, the rules for the structure of phrases appeared. They are a set of grammar rules by which a sentence is constructed. In English, it is formed with the help of a nominal and a verb group. Consider the sentence, "Kate ate the apple". Here, "Kate" is a noun phrase and "ate the apple" is a verb phrase. Different sentences are formed using different structures. As the number of phrase structure rules increases, a parse tree can be created to classify each word in a particular sentence and arrive at its general meaning (Fig. 1).

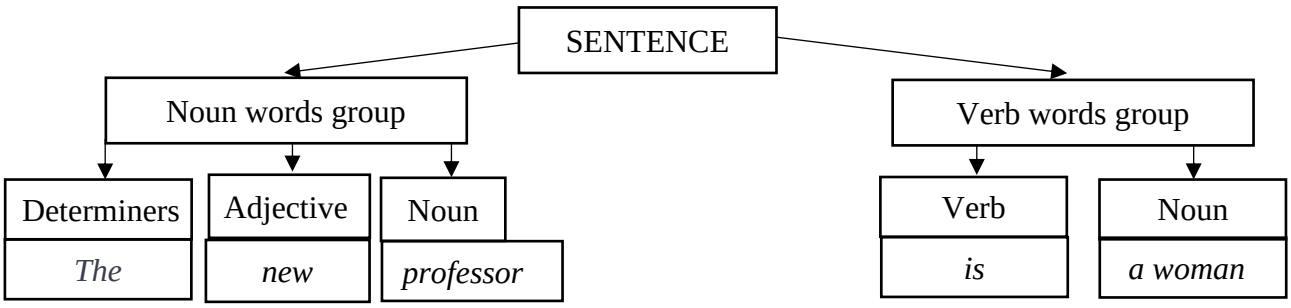

Fig. 1. Parse Tree

The spoken language system integrating speech recognition and speech synthesis is the core technology of human-computer interaction, and oral understanding is the core of spoken language system [5].

Artificial neural networks, created in the form of computer models, cope successfully with the tasks of pattern recognition. They are trainable, they can be easily adapted (and this has already been done) to solve many

practical problems related to speech recognition, control of various machines and devices, event prediction, etc. The biggest feature of a deep neural network is training a large amount of data, and then extracting characteristic information. Characteristic information obtained through a network structure can give good results in the speech comprehension tasks.

An advantage of a neural network in speech processing is that the perceptron can perform discriminant learning between speech units that represent the output classes of the perceptron. The perceptron not only learns and optimizes the parameters for each class on the data belonging to it, but also tries to reject the data belonging to other classes. The perceptron is a structure with a high degree of parallelism, which allows the use of parallel hardware. The first neural network models used in speech recognition systems were developed for static signals, and not for their sequences or signals subject to temporal variability. Later, recurrent neural networks and convolutional neural networks were proposed [5].

As language models are trained on larger and larger texts, the number of unique words (the vocabulary) increases, and continuous space embedding helps to alleviate the curse of dimensionality in language modeling. An alternate description is that a neural net approximates the language function. The neural net architecture might be fed forward or recurrent, and while the former is simpler, the latter is more common. Improved algorithms, powerful computers, and an increase in digitized data have fueled a revolution in machine learning. And new techniques in the 2010s resulted in "rapid improvement in tasks" including language manipulation, in particular, transformer – architecture based on a deep learning model. It was first introduced in 2017 [6]. Table 1 presents the main types of currently existing language models of neural networks.

Table 1

Main types of neural network language models

Language model	Characteristics
BERT-base (2018)	Bidirectional Encoder Representations from Transformers is a new method of pretraining language. BERT is different because it is designed to read in both directions at once. Using this bidirectional capability, BERT is pretrained on two different, but related, NLP tasks: Masked Language Modeling and Next Sentence Prediction [7, 8].
ELMo (2018)	Embeddings from Language Model is a word embedding method for representing a sequence of words as a corresponding sequence of vectors, but unlike BERT, the word embeddings produced by the "bag-of-words" model is a simplifying representation. ELMo embeddings are context-sensitive, producing different representations for words that share the same spelling but have different meanings (homonyms) [9].
GPT (2018)	GPT is a Transformer-based architecture and training procedure for natural language processing tasks. Training follows a two-stage procedure. First, a language modeling objective is used on the unlabeled data to learn the initial parameters of a neural network model. Subsequently, these parameters are adapted to a target task using the corresponding supervised objective [10].
ESPnet (2018)	ESPnet mainly focuses on end-to-end automatic speech recognition (ASR), and adopts widely-used dynamic neural network toolkits, Chainer and PyTorch, as a main deep learning engine. ESPnet also follows the Kaldi ASR toolkit style for data processing, feature extraction/format, and recipes to provide a complete setup for speech recognition and other speech processing experiments [11].
Jasper (2019)	Model uses only 1D convolutions, batch normalization, ReLU, dropout, and residual connections [12].
GPT –2 (2019)	GPT-2 translates text, answers questions, summarizes passages, and generates text output on a level that, while sometimes indistinguishable from that of humans, can become repetitive or nonsensical when generating long passages [13].
WAV2LETTER++ (2019)	It is an open-source deep learning speech recognition framework. wav2letter++ is written entirely in C++, and uses the ArrayFire tensor library for maximum efficiency [14].

Language model	Characteristics
WAV2VEC (2019)	wav2vec, is a convolutional neural network that takes raw audio as input and computes a general representation that can be input to a speech recognition system [15].
XLM (2019)	These are cross-lingual language models (XLMs): one unsupervised that only relies on monolingual data, and one supervised that leverages parallel data with a new cross-lingual language model objective. It obtains state-of-the-art results on cross-lingual classification, unsupervised and supervised machine translation [16].
XLNet (2019)	XLNet uses a generalized autoregressive retraining method that enables learning bidirectional contexts through maximizing the expected likelihood over all permutations of the factorization order and autoregressive formulation. XLNet integrates ideas from Transformer-XL, the state-of-the-art autoregressive model, into retraining [17].
RoBERTa (2019)	This implementation is the same as BERT Model with a tiny embeddings tweak as well as a setup for RoBERTa pretrained models. RoBERTa has the same architecture as BERT, but uses a byte-level BPE as a tokenizer (same as GPT-2) and applies a different pretraining scheme [18].
ELECTRA (2020)	Efficiently Learning Encoder That Accurately Classifies Token Replacements is a new pre-learning method that outperforms development estimation without increasing the computational cost [19].
STC System (2020)	STC system aims at multi-microphone multi-speaker speech recognition and diarization. The system utilizes soft-activity based on Guided Source Separation (GSS) front-end and a combination of advanced acoustic modeling techniques, including GSS-based training data augmentation, multi-stride and multi-stream self-attention layers, statistics layer and spectral [20].
GPT – 3 (2020)	Unlike other models created to solve specific language problems, their API can solve "any problems in English". The algorithm works on the principle of autocompletion: you enter the beginning of the text, and the program generates the most likely continuation of it [21].
ALBERT (2020)	ALBERT incorporates two parameter reduction techniques that lift the major obstacles in scaling pretrained models. The first one is a factorized embedding parameterization. By splitting a large vocabulary embedding matrix into two small matrices, it separates the size of the hidden layers from the size of vocabulary embedding. The second technique is cross-layer parameter sharing. This technique prevents the parameter from growing with the depth of the network [22].
BERT-wwm-ext, 2021)	Pretrained BERT with Whole Word Masking due to the complexity of Chinese grammar structure and the semantic diversity, a BERT (wwm-ext) was proposed based on the whole Chinese word masking, which mitigates the drawbacks of masking partial Word Piece tokens in pretrained BERT [23].
PaLM (2022)	This is Pathways Language Model 540-billion parameter, dense decoder. Only Transformer model trained with the Pathways system enabled us to efficiently train a single model across multiple TPU v4 Pods [24].

As can be seen from Table 1, the first transformer models, using a bidirectional capability, allowed two different but related tasks of the NLP to be studied beforehand: simulating a masked language and predicting the next sentence. Bidirectional Encoder Representations from Transformers consist of two steps: the first step is pretraining where the data enter the layers of transformer, and the result of this step are vectors for words. The second step is fine tuning. The pretraining step consists of two steps: the masked LM and Next Sentence Prediction (NSP) [7, 8]. BERT is not without flaws, the most obvious one is the learning method – the neural network tries to guess each word separately, which means that it loses some possible connections between words during the learning process. Another one is that the neural network is trained on masked tokens, and then used fundamentally different tasks, more complex ones.

Embeddings from Language Model is a deep contextualized word representation that models both complex characteristics of word usage (e.g., syntax and semantics), and how this usage varies across linguistic contexts (i.e., to model polysemy), such as "bank" in "river bank" and "bank balance". These word vectors are learned functions of the

internal states of a deep bidirectional language model (biLM), which is pretrained on a large text corpus. They can be easily added to existing models and significantly improve the state of the art across a broad range of challenging NLP problems, including question answering, textual entailment, and sentiment analysis [9].

To alleviate the problem, suffering from the discrepancy between the pretraining and fine-tuning stage because the masking token [MASK] never appears on the fine-tuning stage, XLNet was proposed, which is based on Transformer-XL. To achieve this goal, a novel two-stream self-attention mechanism, and one to change the autoencoding language model into an autoregressive one, which is similar to the traditional statistical language models, were proposed [17]. RoBERTa, STC System, GPT models were used in quite a large number of systems. And they showed pretty good results. These models suggested that averaging all token representations consistently induced better sentence representations than using the token embedding; combining the embeddings of the bottom layer and the top layer outperformed the use of the top two layers; and normalizing sentence embeddings with a whitening algorithm consistently boosted the performance [18, 20, 21].

The next step, probably, will be to study the oversampling and undersampling of textual data to improve the overall entity recognition effect.

Results. The analysis of the literary sources describing new methods of processing oral speech, which provides information about neural networks, methods for the structure and synthesis of speech, made it possible to detect the following:

1. All the models presented in the review require large computing power to solve natural language processing problems. It is computationally more expensive due to its larger structure.

2. None of the currently existing technologies enable solving the full range of tasks for recognizing continuous, defective speech.

3. Most natural language processing models are designed to handle a wide variety of English dialects and idioms.

Discussion and Conclusions. Voice assistants reproduce and reinforce all stereotypes algorithms. They, as a rule, reproduce those stereotypes that exist now in society. What does this achievement really mean? It means that the voice assistant is no worse (or maybe even better) than an average person at recognizing the speech of a person with a standard North American accent. But if an African American speaks to an assistant, then the accuracy will drop to about 80 %. This is a huge difference. Moreover, when converting voice to text, the specifics of writing, which can be important for speakers, are guaranteed to be lost.

Voice assistants do not take into account the speech and user habits of the elderly and people with special needs.

And here, it is not even always the complexity of recognition. There is, e.g., such a condition as dysarthria – a feature of the functioning of the connections between the speech apparatus and the nervous system, which can cause difficulties in pronouncing individual sounds or, in general, in speech.

Also, due to hardware limitations, any cartridge will result in too many model parameters and unsuccessful execution. The way to solve the problem of multiple cycles of dialogue requires further research.

References

1. Lee A, Auli M, Ranzato MA. Discriminative reranking for neural machine translation. In: ACL-IJCNLP 2021 – 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Proceedings of the Conference, 2021. P. 7250–7264. http://dx.doi.org/10.18653/v1/2021.acl-long.563

2. Prashant Johri, Sunil Kumar Khatri, Ahmad T Al-Taani, et al. Natural language processing: History, evolution, application, and future work. In: Proc. 3rd International Conference on Computing Informatics and Networks. 2021;167:365–375. http://dx.doi.org/10.1007/978-981-15-9712-1_31

3. Nitschke R. Restoring the Sister: Reconstructing a Lexicon from Sister Languages using Neural Machine Translation. In: Proc. 1st Workshop on Natural Language Processing for Indigenous Languages of the Americas, AmericasNLP 2021. 2021. P. 122–130. http://dx.doi.org/10.18653/v1/2021.americasnlp-1.13

4. Pokrovskii MM. Izbrannye raboty po yazykoznaniyu. Moscow : Izd-vo Akad. nauk SSSR; 1959. 382 p. (In Russ.)

5. Ryazanov VV. Modeli, metody, algoritmy i arkhitektury sistem raspoznavaniya rechi. Moscow: Vychislitel'nyi tsentr im. A.A. Dorodnitsyna; 2006. 138 p. (In Russ.)

6. Lixian Hou, Yanling Li, Chengcheng Li, et al. Review of research on task-oriented spoken language understanding. Journal of Physics Conference Series. 2019;1267:012023. http://dx.doi.org/10.1088/1742-6596/1267/1/012023

7. Ashish Vaswani, Noam Shazeer, Niki Parmar, et al. Attention Is All You Need. In: proc. 31st Conf. on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA. https://doi.org/10.48550/arXiv.1706.03762

8. Jacob Devlin, Ming-Wei Chang, Kenton Lee, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Computing Research Repository. 2018. P. 1−16.

9. Matthew E Peters, Mark Neumann, Mohit Iyyer, et al. Deep contextualized word representations. In: Proc. NAACL-HLT. 2018;1:2227–2237.

10. Alec Radford, Karthik Narasimhan, Tim Salimans, et al. Improving Language Understanding by Generative Pre-Training. Preprint. https://pdf4pro.com/amp/view/improving-language-understanding-by-generative-pre-training-5b6487.html

11. Shinji Watanabe, Takaaki Hori, Shigeki Karita, et al. ESPnet: End-to-End Speech Processing Toolkit. 2018. https://arxiv.org/abs/1804.00015

12. Jason Li, Vitaly Lavrukhin, Boris Ginsburg, et al. Jasper: An End-to-End Convolutional Neural Acoustic Model. 2019. https://arxiv.org/abs/1904.03288

13. Chaitra Hegde, Shrikumar Patil. Unsupervised Paraphrase Generation using Pre-trained Language Models. 2020. https://arxiv.org/abs/2006.05477

14. Vineel Pratap, Awni Hannun, Qiantong Xu, et al. Wav2letter++: The Fastest Open-source Speech Recognition System. In: Proc. 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). https://doi.org/10.1109/ICASSP.2019.8683535

15. Schneider S, Baevski A, Collobert R, et al. Wav2vec: Unsupervised Pre-Training for Speech Recognition. In: Proc. Interspeech 2019, 20th Annual Conference of the International Speech Communication Association. P. 3465−3469. https://doi.org/10.21437/Interspeech.2019-1873

16. Alexis Conneau, Guillaume Lample. Cross-lingual Language Model Pretraining. In: Proc. 33rd Conference on Neural Information Processing Systems. 2019. P. 7057−7067.

17. Zhilin Yang, Zihang Dai, Yiming Yang, et al. XLNet: Generalized Autoregressive Pretraining for Language Understanding. 2019. https://arxiv.org/abs/1906.08237

18. Yinhan Liu, Myle Ott, Naman Goyal, et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach. ICLR 2020 Conference Blind Submission. 2019. https://doi.org/10.48550/arXiv.1907.11692

19. Manning Kevin Clark, Minh-Thang Luong, Quoc V Le, et al. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. ICLR 2020 Conference Blind Submission. 2020. https://openreview.net/forum?id=r1xMH1BtvB

20. Medennikov I, Korenevsky M, Prisyach T, et al. The STC System for the CHiME-6 Challenge. In: Proc. 6th International Workshop on Speech Processing in Everyday Environments (CHiME 2020). 2020. P. 36−41.

21. Greg Brockman, Mira Murati, Peter Welinder. OpenAI API. 2020 : OpenAI Blog. https://openai.com/blog/openai-api/

22. Zhenzhong Lan, Mingda Chen, Sebastian Goodman, et al. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. 2020. https://arxiv.org/abs/1909.11942

23. Yiming Cui, Wanziang Che, Ting Liu, et al. Pre-Training With Whole Word Masking for Chinese BERT. In: IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2021;29:3504−3514. https://doi.org/10.1109/TASLP.2021.3124365

24. Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, et al. PaLM: Scaling Language Modeling with Pathways. 2022. https://arxiv.org/abs/2204.02311

Received 10.04.2022
Revised 06.05.2022
Accepted 06.05.2022

About the Authors:

Kazakova, Maria A., graduate student of the Applied Mathematics and Computer Science Department, Kazan National Research Technical University named after A.N. Tupolev–KAI (10, K. Marx St., Kazan, 420111, RF), ResearcherID, ORCID, kazakovamaria2609@gmail.com

Sultanova, Alina P., associate professor of the Department of Foreign Languages, Russian, Russian as a Foreign Language, Kazan National Research Technical University named after A.N. Tupolev–KAI (10, K. Marx St., Kazan, 420111, RF), ResearcherID, ORCID, alinasultanova@mail.ru

Claimed contributorship

M. A. Kazakova: basic concept formulation; goals and objectives of the literary review; search and analysis of literary sources; text preparation. A. P. Sultanova: generalization of the conceptual provisions of philology as a basis for the development of oral speech processing methods; the text revision.

Conflict of interest statement

The authors do not have any conflict of interest.

All authors have read and approved the final manuscript.